ADVERSARIAL SAMPLE DETECTION FOR SPEAKER VERIFICATION BY NEURAL VOCODERS

Haibin Wu[1], *Po-chun Hsu*[1], *Ji Gao*[2], *Shanshan Zhang*[2], *Shen Huang*[2], *Jian Kang*[2],
Zhiyong Wu[3], *Helen Meng*[4], *Hung-yi Lee*[1]

[1] Graduate Institute of Communication Engineering, National Taiwan University
[4] Centre for Perceptual and Interactive Intelligence, The Chinese University of Hong Kong
[3] Shenzhen International Graduate School, Tsinghua University
[2] Tencent Research, Beijing, China

OUTLINE

- **Motivation**
- **Background**
- **Proposed Method**
- **Experiment**
- **Conclusion**

1. Motivation

- Automatic speaker verification (ASV), one of the most important technology for biometric identification, has been widely adopted in security-critical applications.

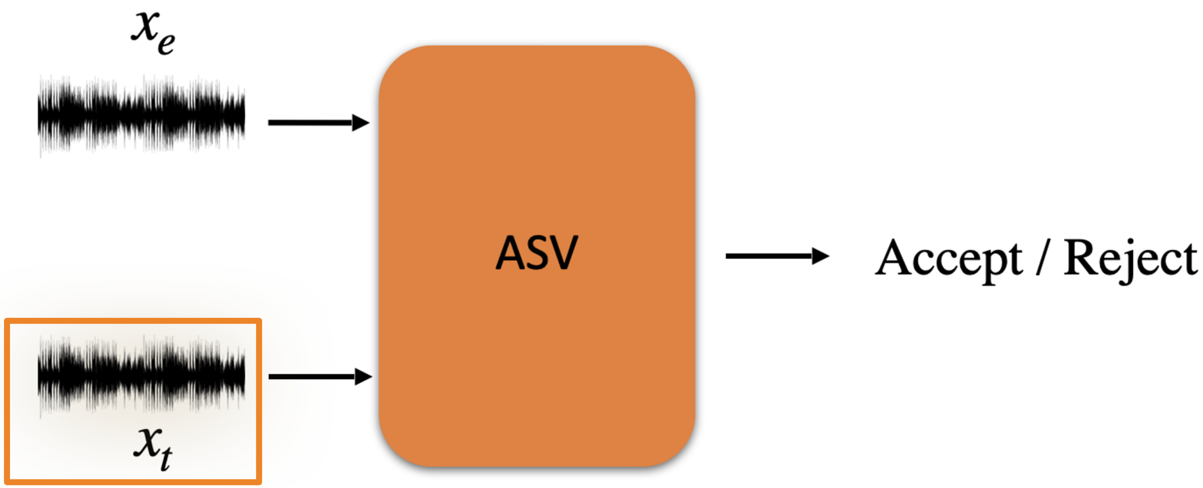
- ASV is seriously vulnerable to recently emerged adversarial attacks, yet effective countermeasures against them are limited.

2. Background

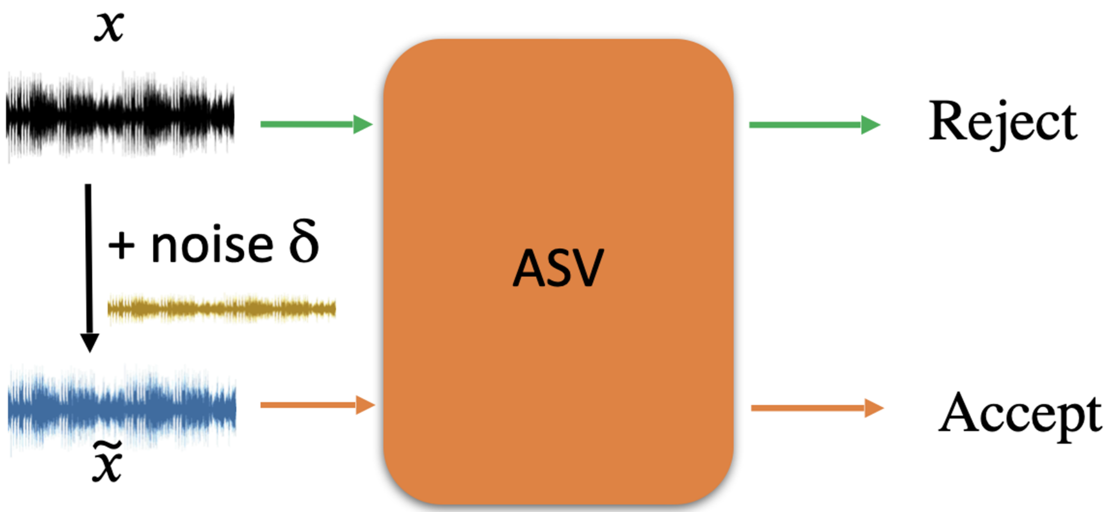
2.1 Automatic speaker verification

2.2 Adversarial attack

2.1 Automatic speaker verification



2.2 Adversarial attack



3. Proposed Method

3.1 Implementation

3.2 Rationales

3.1 Implementation

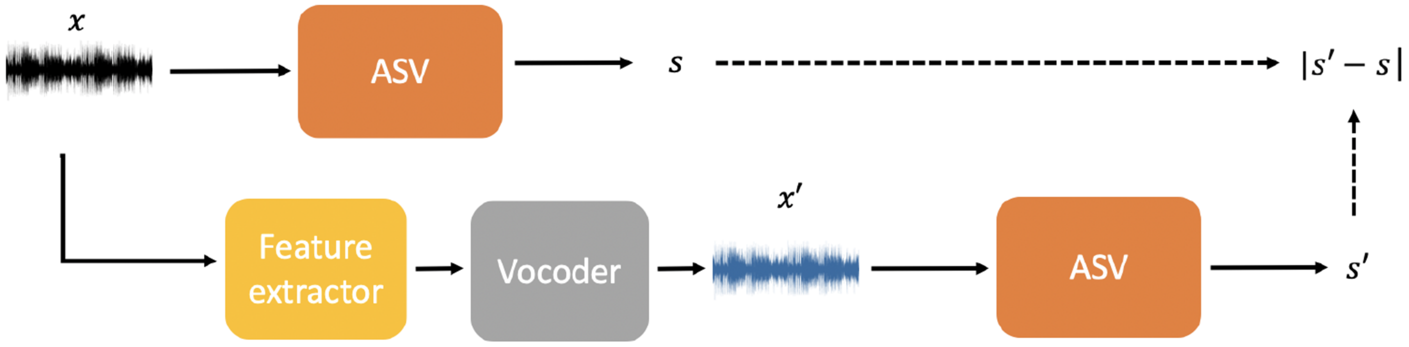
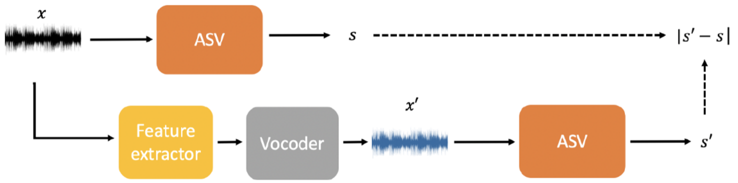


Fig. 1. Proposed detection framework. s and s' are the ASV scores for x and x'. $|s - s'|$ is the absolute value between s and s'.

3.1 Implementation



$$\mathbb{T}_{gen} = \{x_{gen}^1, x_{gen}^2, \ldots, x_{gen}^I\} \qquad d = |s - s'|$$

$$d_{gen}^i = |s_{gen}^i - s_{gen}^i{}'|, for \;\; i = 1, 2..., I$$

$$FPR_{det}(\tau) = \frac{|\{d_{gen}^i > \tau : x_{gen}^i \in \mathbb{T}_{gen}\}|}{|\mathbb{T}_{gen}|}$$

$$\tau_{det} = \{\tau \in \mathbb{R} : FPR_{det}(\tau) = FPR_{given}\}$$

3.1 Implementation

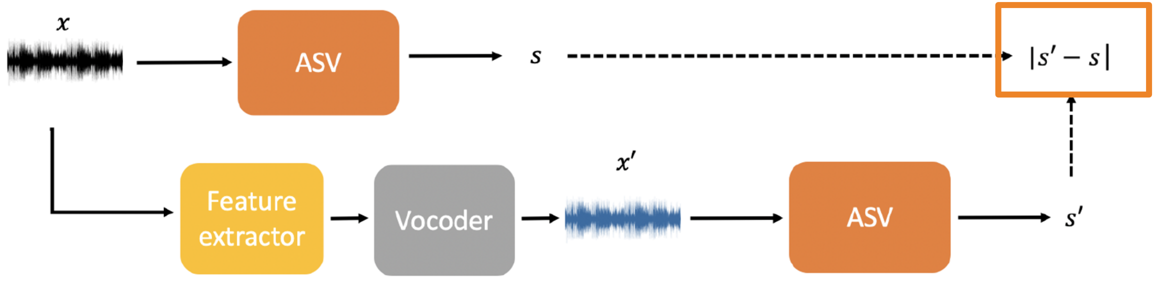


Fig. 1. Proposed detection framework. s and s' are the ASV scores for x and x'. $|s - s'|$ is the absolute value between s and s'.

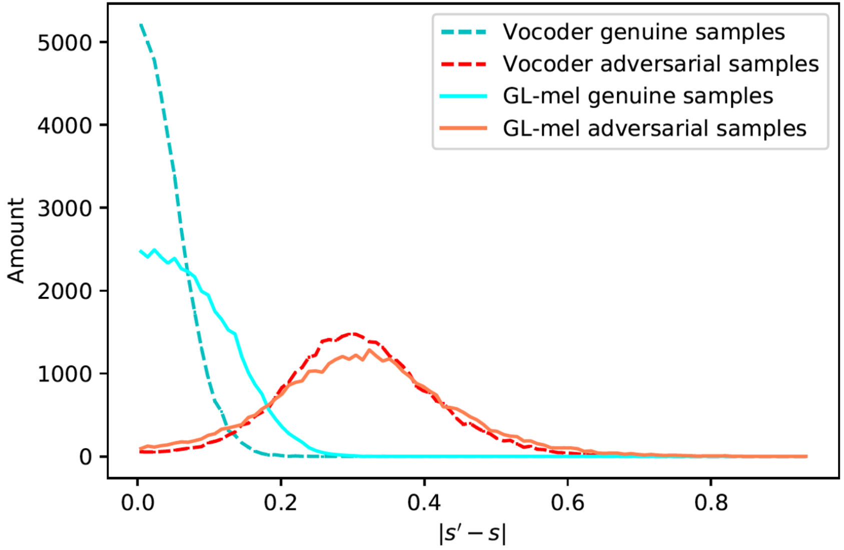
$$FPR_{det}(\tau) = \frac{|\{d^i_{gen} > \tau : x^i_{gen} \in \mathbb{T}_{gen}\}|}{|\mathbb{T}_{gen}|}$$

$$\boxed{\tau_{det}} = \{\tau \in \mathbb{R} : FPR_{det}(\tau) = FPR_{given}\}$$

3.2 Rationales

- As the vocoder is data-driven and trained with genuine data during training, it models the distribution of genuine data, resulting in less distortion when re-generating genuine waveforms.
- Thus, during inference, the vocoder‘s preprocessing will not influence the ASV scores of genuine samples too much.
- However, suppose the inputs are adversarial samples. In that case, the vocoder will try to pull it back towards the manifold of their genuine counterparts to some extent, resulting in purifying the adversarial noise.

3.2 Rationals



- The score difference for genuine samples is near zero.
- While the score difference for adversarial samples is much larger.
- We can simply set a threshold value to distinguish them.

4. Experiment

4.1 Experimental setup

4.2 Experimental result

4.1 Experimental setup

- The ResNet backbone is trained by Voxceleb2 as the speaker embedding extractor.
- The Basic iterative method is used for crafting the adversarial sampels.
- We use a traditional vocoder, the Griffin-Lim and a neural vocoder, ParallelWaveGAN for detection.

4.2 Experimental results

Table 1. EER with different ϵ

Method	EER with different ϵ (%)				
	20	15	10	5	0 (no attack)
None	99.33	95.66	90.57	74.04	2.88
Vocoder	87.58	65.75	52.20	30.37	3.39
GL-lin	95.23	80.83	66.73	39.49	3.93
GL-mel	88.41	65.39	49.76	26.67	3.81

4.2 Experimental results

Table 1. EER with different ϵ

Method	EER with different ϵ (%)				
	20	15	10	5	0 (no attack)
None	99.33	95.66	90.57	74.04	2.88
Vocoder	87.58	65.75	52.20	30.37	3.39
GL-lin	95.23	80.83	66.73	39.49	3.93
GL-mel	88.41	65.39	49.76	26.67	3.81

- When testing on genuine samples, the EER is 2.88%. When using generated speech as inputs, the EER slightly increased.

4.2 Experimental results

Table 1. EER with different ϵ

Method	EER with different ϵ (%)				
	20	15	10	5	0 (no attack)
None	99.33	95.66	90.57	74.04	2.88
Vocoder	87.58	65.75	52.20	30.37	3.39
GL-lin	95.23	80.83	66.73	39.49	3.93
GL-mel	88.41	65.39	49.76	26.67	3.81

- While introducing the adversarial attack, the EER increased from 2.88% to over 70%, which shows the effectiveness of the attack method.

4.2 Experimental results

Table 1. EER with different ϵ

Method	EER with different ϵ (%) 20	15	10	5	0 (no attack)
None	99.33	95.66	90.57	74.04	2.88
Vocoder	87.58	65.75	52.20	30.37	3.39
GL-lin	95.23	80.83	66.73	39.49	3.93
GL-mel	88.41	65.39	49.76	26.67	3.81

- The vocoder has slight purification performance.
- However, the re-synthesis process will not affect the genuine EER too much.

4.2 Experimental results

Table 3. Detection rate with different ϵ

FPR_{given}	Method	Detection rate with different ϵ (%) 20	15	10	5
0.05	Vocoder	**99.76**	**98.82**	**97.30**	**89.33**
	Vocoder-L	99.38	97.23	94.07	81.21
	GL-lin	89.12	88.30	84.64	71.29
	GL-mel	95.39	91.33	85.37	68.07
	Gaussian	34.54	51.29	61.56	68.57
0.01	Vocoder	**98.92**	**97.56**	**94.76**	**81.60**
	Vocoder-L	97.96	94.37	88.77	70.15
	GL-lin	73.62	73.63	70.62	56.37
	GL-mel	87.98	82.27	75.04	56.07
0.005	Vocoder	**98.30**	**96.78**	**93.25**	**78.21**
	Vocoder-L	96.78	92.58	85.81	64.65
	GL-lin	64.76	64.97	62.85	49.32
	GL-mel	83.94	77.71	70.47	51.42
0.001	Vocoder	**96.04**	**93.89**	**88.60**	**68.58**
	Vocoder-L	93.36	87.34	78.24	53.18
	GL-lin	45.10	45.27	44.72	34.28
	GL-mel	72.53	65.98	59.66	40.98

- We find that using Vocoder performs the best among all methods. In most cases, more than 90% of the adversarial samples could be detected.
- For Griffin-Lim based methods, we find that they might be good approaches for detection with a large FPR. However, in stricter cases, the detection rates decrease drastically.

4.2 Experimental results

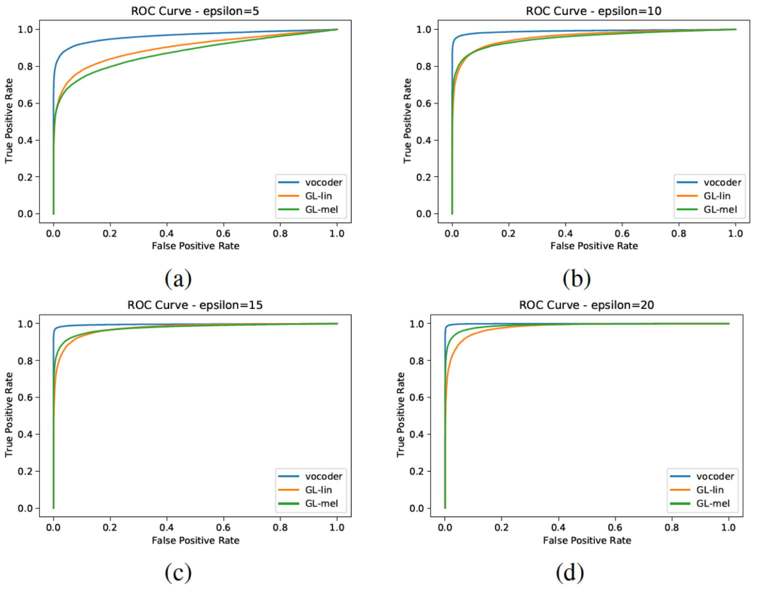


Fig. 3. ROC curve under different epsilon (ϵ)

- The larger the area under the curve (AUC) is, the better the detection performance.
- The vocoder based detection method attains very high AUC, almost near 1.

5. Conclusion

- This work adopts the neural vocoder to detect adversarial samples for ASV.
- The proposed method achieves effective detection performance.
- For the future work, we will evaluate the detection performance when the detection method is known to the attackers.

THANK YOU!